计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140808
52(11):2451-2459,2015

基于深度神经网络的有色金属领域实体识别

毛存礼[1] 余正涛[1] 沈 韬[2] 高盛祥[1] 郭剑毅[1] 线岩团[1]
[1](昆明理工大学信息工程与自动化学院 昆明 650500)
[2](昆明理工大学材料科学与工程学院 昆明 650093)
(maocunli@163.com)

A Kind of Nonferrous Metal Industry Entity Recognition Model Based on Deep Neural Network Architecture

Mao Cunli[1], Yu Zhengtao[1], Shen Tao[2], Gao Shengxiang[1], Guo Jianyi[1], and Xian Yantuan[1]
[1](*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500)
[2](*Faculty of Material Science and Engineering, Kunming University of Science and Technology, Kunming* 650093)

Abstract Aimed at entity recognition in the field of nonferrous metal, and oriented the complex and strongly nested structure characteristics of domain entity such as product names, organizations, and placenames, it is proposed a kind of nonferrous metal industry entity recognition model based on deep neural network (DNN) architecture. In order to effectively use the characteristics of tight combination between the characters of domain entity and to bypass the Chinese words segmentation in the professional field, the model uses neural networks to automatically learn the word embeddings vector representation of Chinese characteristics as its inputs. The denoising autoencoder (DAE) of text window makes some pre-training on each DNN hidden layer. The pre-training extracts optimal feature vector combination which will be used in nonferrous metal domain entity recognition. Moreover, we detailedly describe the pre-training process that the denoising autoencoder of text window based on neuron language model makes and the construction process of deep network on nonferrous metal entity recognition. Finally, to validate the method's effectiveness, we make some experiments of entity recognition on several entity types, such as product names, mineral names and place names in nonferrous metal domain. The experimental results show that the proposed model has good effect on the entity recognition of the professional domain.

Key words field of nonferrous metal; deep neural network(DNN); word embeddings; denoising autoencoder(DAE); entity recognition

摘 要 针对有色金属领域实体识别问题,提出一种基于深度神经网络(deep neural network, DNN)架构的有色金属领域实体识别方法.为能有效获取有色金属领域实体中字符间的紧密结合特征,并回避专业领域中文分词问题,使用神经网络的方法自动学习中文字符 embeddings 向量化表示作为模型输入.基于降噪自动编码器(denoising autoencoder, DAE)对深度神经网络的每个隐层进行逐层预训练获取用于有色金属领域实体识别的最优特征向量组合,并详细介绍了基于神经语言模型的文本窗口降噪自动

收稿日期:2014-09-03;**修回日期**:2015-06-02
基金项目:国家自然科学基金项目(61262041,61472168,61163004);"科技部创新人才推进计划"中青年科技创新领军人才配套项目(2014HE001);云南省自然科学基金重点项目(2013FA030)
通信作者:余正涛(ztyu@hotmail.com)

编码器预训练及有色金属实体识别的深层网络构建过程.为验证方法的有效性,对有色金属领域产品名、矿产名、地名、组织机构4类实体识别进行实验.实验结果表明,提出的方法对于专业领域的实体识别具有较好的效果.

关键词 有色金属领域;深度神经网络;词汇 embeddings;降噪自动编码器;实体识别

中图法分类号 TP391

准确识别有色金属领域文本中的实体是对有色金属领域文本进行深层次分析的基础.目前,实体识别方法大致分为2类:1)基于规则的方法,如Mikheev等人[1]使用人工构建的规则抽取预先定义的各种类别的实体;2)基于统计机器学习的命名实体识别方法成为当前的主流方法,包括:隐马尔可夫模型(hidden Markov models, HMMs)、条件随机场(conditional random fields, CRFs)模型、最大熵(maximum entropy, ME)模型等,大量的研究工作主要集中在针对英文领域的实体识别任务[2-4].在面向特定领域的实体识别方面也开展了一些研究,如针对生物医学领域实体识别[5]、针对Tweets短文本的实体识别方法[6-7]、针对商务领域产品命名实体识别[8]等.以上方法都是基于浅层模型的统计机器学习方法,这类方法在特征选择过程中过多依赖于人工参与,并且在模型训练前期需要人工标注大量的训练语料.为避免传统浅层学习模型在特征选择方面的不足,基于深度学习的特征提取方法成为当前机器学习的研究热点,深度学习模型在选择特征时,模型能通过多个层次的逐层训练将原始特征进行多次非线性变换,自动获取到最稳定的特征用于模型训练.近年来,深度学习在语音、图像方面取得显著的成效[9-12],在自然语言处理任务中也得到了很好的应用,如Collobert等人[13]提出一种基于神经语言模型的深度网络模型,用于词性标注、语块分析、实体识别、语义角色标注的任务时能取到较好的效果.Chen等人[14]提出利用深信任网络(deep belief nets, DBN)模型进行基于特征的实体关系抽取,有效解决了基于高维空间特征的信息抽取问题.

对于有色金属领域实体识别任务,不仅实体名称组成复杂、结构嵌套,还面临领域分词困难以及缺乏大规模人工标注训练语料等诸多问题.针对这些问题,本文提出一种基于深度学习的有色金属实体识别模型,该模型首先将输入的中文字符映射成embeddings向量,并将这些向量拼接成一个初始的组合特征向量作为深度神经网络(deep neural network, DNN)模型的输入,由DNN模型的多个隐层(文本窗口降噪自动编码器)的逐层预训练自动提取到最优的特征向量组合,用于训练有色金属实体分类器.实验结果表明,我们提出的方法相比CRFs及BP(back propagation)神经网络模型具有较好的效果.

1 有色金属实体识别DNN架构

我们将有色金属领域实体识别任务当作序列化标注问题,为每个中文字符分配一个唯一的标签,提出的有色金属领域实体识别深层架构如图1所示:

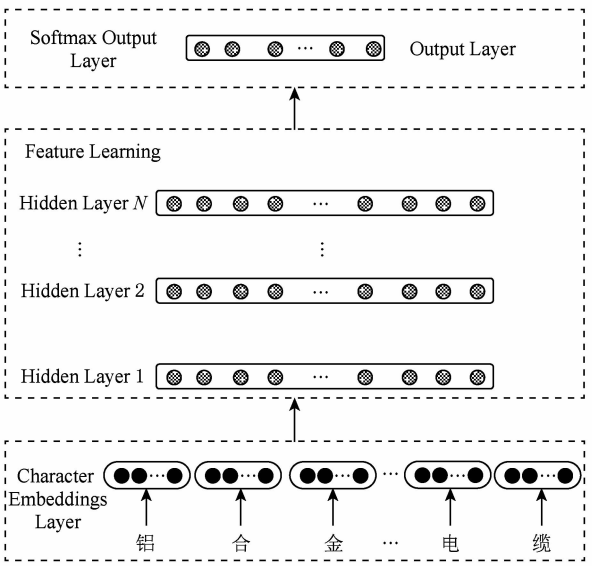


Fig. 1 DNN architecture of entity recognition in the field of nonferrous metal.

图1 有色金属实体识别DNN架构

图1中,模型的输入是一个固定大小的文本窗口,我们将一个句子包含的中文字符当作一个窗口,每个文本窗口只有一个中心字符;模型的输出是计算文本窗口中心字符所有可能标记(如B-M_PRO, I-M_PRO, B-M_ORG, I-M_ORG等)的得分.本文针对有色金属实体识别任务采用IOBES标记方案[13],即:B-M_PRO标记有色金属产品实体第1个字符,I-M_PRO标记有色金属产品内部字符,E-M_PRO标记有色金属产品实体最后1个字符,S-M_PRO

标记单个字符独自构成有色金属产品实体,O-M_PRO 标记不是有色金属产品实体中的字符.词向量层(character embeddings layer)执行矩阵查表(lookup table)操作,提取输入文本窗口中的每个字符的 embeddings 向量,并将这些特征向量拼接起来作为网络中第一个隐层的输入;输出层是 1 个 softmax 层,输出每 1 个节点相应的标记及概率.

2 基于 Word Embeddings 的词汇向量化表示

传统的词向量表示方法,通常将词汇表示为 one hot 向量形式,该方法是将每个词表征成长度为 N(语料中词典大小)的向量,其中绝大多数元素为 0,只有一个维度的值为 1,这个维度就代表当前的词.然而,这种词向量表征方法存在的问题是出现严重的数据稀疏,并且丢失了词语间的语义关系.深度学习在自然语言处理方面的优势还在于能够对词汇进行更深层次抽象,将词汇表征成一个高密度的低维实数向量,能有效避免传统 one hot 形式的向量表示方法出现的数据稀疏问题,并且能很好地表征词语之间的词法、句法和语义信息.如 Collobert 等人[13]提出的基于神经语言模型自动学习词汇的 word embeddings 向量化表示方法,其核心思想是一个词包含的语义应该由该词周围的词决定,该方法能够将适合出现在窗口中间位置的词聚合在一起,而将不适合出现在这个位置的词分离开来,从而将语法或词性相似的词映射到向量空间中相近的位置.Mikolov 等人[15]提出了一种使用周围词来预测中间词的连续词袋(continuous bag-of-words, CBOW)模型,该模型的核心思想是将输入层相邻的词向量直接相加得到隐层,并用隐层预测中间词的概率,这种方法的优点是周围词的位置不会影响到预测的结果.Mikolov 等人[15]还提出了一种连续 skip-gram 模型来表征词向量,该模型与 CBOW 的预测方式正好相反,是通过中间词来预测周围词的概率.

相比普通实体结构,有色金属领域实体具有组成复杂、结构嵌套及较强的领域特点,如"华伦铝合金电缆".为获取有色金属领域实体内部字符之间的语法结构特征,回避领域分词对实体识别结果的影响,本文采用 Collobert 等人提出的词向量表示方法,将每一个中文字符进行向量化表示作为模型初始化输入.通过执行矩阵 Lookup Table 操作,将文本窗口中输入的每个中文字符转换成对应的 embeddings 向量,如图 2 所示.对于文本窗口中的每个中文字符 $c\in D$,1 个 d_{wrd} 维的内部特征向量的表征是通过 Lookup Table 层的 $L_w(\cdot)$ 操作得到 $L_w(c)=\langle \boldsymbol{W}\rangle_c^1$.这里 $\boldsymbol{W}\in\mathbb{R}^{d_{\text{wrd}}\times|D|}$ 是一个要学习的参数矩阵,$\langle \boldsymbol{W}\rangle_c^1\in\mathbb{R}^{d_{\text{wrd}}}$ 是矩阵 $\boldsymbol{W}$ 中字符 c 对应的列,d_{wrd} 是中文字符 embeddings 向量的大小(是由用户选择的超参数,通常为 50 或 100).

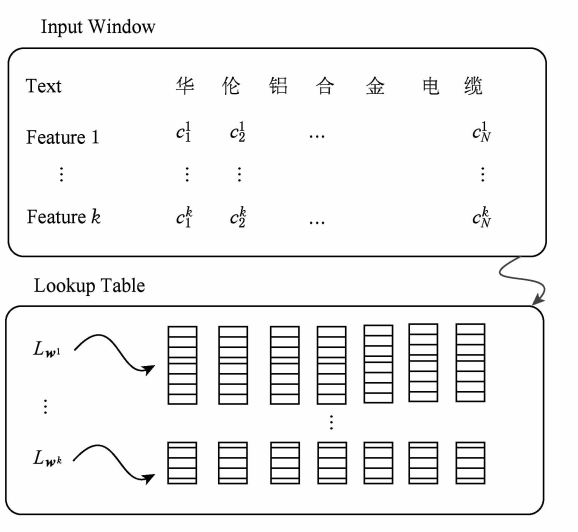


Fig. 2 Embeddings vector of Chinese characters.

图 2 中文字符向量化表示

对于给定的一个文本窗口 T 的任意一个序列 $[c]_1^T\in D$,Lookup Table 层对文本窗口中的每个中文字符执行相同的操作,产生输入矩阵:

$$L_w([c]_1^T)=(\langle \boldsymbol{W}\rangle_{[c]_1}^1\langle \boldsymbol{W}\rangle_{[c]_2}^1\cdots\langle \boldsymbol{W}\rangle_{[c]_T}^1). \quad (1)$$

我们把一个中文字符表征成 k 个离散特征 $c\in D^1\times D^2\times\cdots\times D^k$,这里 D^k 是中文字符 c 的第 k 个特征词典,我们把每个特征与带有参数 $\boldsymbol{W}^k\in\mathbb{R}^{d_{\text{wrd}}^k\times|D^k|}$ 的查表 $L_{\boldsymbol{W}^k}(\cdot)$ 关联起来,这里 $d_{\text{wrd}}^k\in\mathbb{N}$ 是由用户指定的向量大小.给定一个中文字符 c,该字符对应的一个 $d_{\text{wrd}}=\sum_k d_{\text{wrd}}^k$ 维的特征向量是通过串联所有 Lookup Table 输出得到:

$$L_{\boldsymbol{W}^1,\cdots,\boldsymbol{W}^k}(c)=\begin{pmatrix}L_{\boldsymbol{W}^1}(c_1)\\ \vdots\\ L_{\boldsymbol{W}^k}(c_k)\end{pmatrix}=\begin{pmatrix}\langle \boldsymbol{W}^1\rangle_{c_1}^1\\ \vdots\\ \langle \boldsymbol{W}^k\rangle_{c_k}^1\end{pmatrix}. \quad (2)$$

对于一个中文字符序列 $[c]_1^T$ 的 Lookup Table 层的输出矩阵与式(1)相似,但是每个离散特征增加到额外的行:

$$L_{\boldsymbol{W}^1,\cdots,\boldsymbol{W}^k}([c]_1^T)=\begin{pmatrix}\langle \boldsymbol{W}^1\rangle_{[c_1]_1}^1 & \cdots & \langle \boldsymbol{W}^1\rangle_{[c_1]_T}^1\\ \vdots & & \vdots\\ \langle \boldsymbol{W}^k\rangle_{[c_k]_1}^1 & \cdots & \langle \boldsymbol{W}^k\rangle_{[c_k]_T}^1\end{pmatrix}. \quad (3)$$

由于本文的任务是识别有色金属领域的实体，为此，在映射文本窗口中每个字符通用的 embeddings 特征向量时，需要扩展一些有助于有色金属领域实体识别的离散特征，如将有色金属领域地名库、有色金属领域组织机构库、有色金属领域产品库、有色金属元素库等离散特征扩展到中文字符对应的 embeddings 向量中.

上述的 Lookup Table 中执行的查表操作是假定每个中文字符对应的 embeddings 向量已经生成. 实际上，由于 $\boldsymbol{W}$ 是一个参数矩阵，需要利用神经网络的方法自动学习出与具体任务相关的 embeddings 向量. 这是一个无监督的训练过程，首先将字典中的每一个中文字符随机初始化为一个向量，然后使用大规模无标记的有色金属领域文本数据作为训练语料来优化此向量，从而得到最优的中文字符向量化表示.

3 基于降噪自动编码器的隐层预训练

3.1 神经语言模型文本窗口降噪自动编码器

针对文本窗口中字符序列标注任务(如词性标注、实体识别等)，利用神经语言模型来训练一个文本窗口降噪自动编码器(denoising autoencoder, DAE)模型是一种非常有效的方法[16-18]. 这种训练方法假定窗口中各个词的序列标记主要依赖于与它临近词的标记. 我们提出的基于文本窗口的 DAE 模型由编码器(encoder)和解码器(decoder)构成. encoder 的作用是处理有噪音的输入数据，并生成一个高密度的低维实数向量作为数据的“encoder”特征；decoder 的作用是处理“encoder”，并恢复出清晰的数据. DAE 模型是通过重构误差准则来训练，该训练准则是度量重构后的清晰数据与原始数据之间的误差. 我们忽略了中心字符作为引入的噪声，并不断让模型来重构给定上下文信息的中心字符作为降噪过程.

例如，文本窗口“华伦铝合金电缆”，我们给定这个文本窗口的上下文“华伦铝金电缆”，文本窗口 DAE 模型的任务是让模型自动恢复被省略的中心字符“合”. 处理方法是 DAE 模型通过给每一个字符分配在给定上下文环境下的一个概率来处理这个任务，这些概率被看作是原始中心字符“合”的 one hot 形式向量化表示重构. 完成这个任务的具体处理过程，如图 3 所示.

我们以 1 个单隐层网络为例来介绍神经语言模型[16-17]：

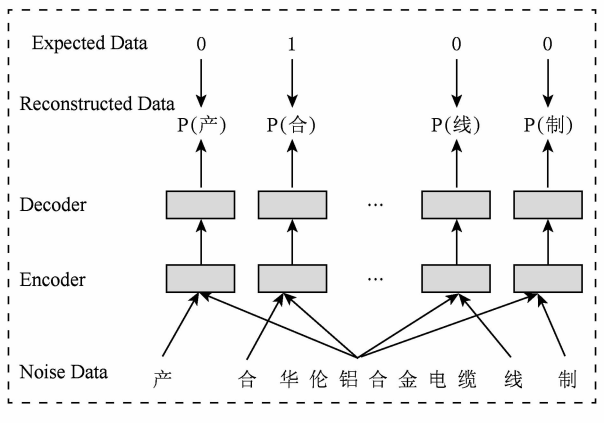


Fig. 3 Example of text window DAE based on neural language model.

图 3 基于神经语言模型的文本窗口 DAE 模型举例

$$l_1(c_1,c_2,\cdots,c_s)=l_{\text{hidden}}(l_{\text{in}}(c_1,c_2,\cdots,c_s))=\tanh(\boldsymbol{M}\cdot(L_w(c_1)),\boldsymbol{M}\cdot(L_w(c_2)),\cdots,\boldsymbol{M}\cdot(L_w(c_s))+b), \tag{4}$$

其中，$\boldsymbol{M}$ 表示 Encoder 层的参数矩阵.

$$l_2(c_1,c_2,\cdots,c_s)=l_{\text{out}}(l_1(c_1,c_2,\cdots,c_s))=\text{sigmoid}(\boldsymbol{M}\cdot(L_w(c_1)),\boldsymbol{M}\cdot(L_w(c_2)),\cdots,\boldsymbol{M}\cdot(L_w(c_s))+b). \tag{5}$$

文本窗口 DAE 模型的 encoder 形式如下：

$$encoder(X)=(l_1(X,c_1),l_1(X,c_2),\cdots,l_1(X,c_s)), \tag{6}$$

其中，$X=(c_{-s/2}^{-1},c_1^{s/2})$，表示省略中心字符的文本窗口，其特征表示为

$$f(X)=(y_1,y_2,\cdots,y_n), \tag{7}$$

其中，$y_i=l_1(X,c_i)$.

$decoder$ 形式为

$$decoder(Y)=(l_2(y_1),l_2(y_2),\cdots,l_2(y_n)). \tag{8}$$

文本窗口 DAE 模型的最小恢复错误形式为

$$\begin{cases}E(\theta,c_{-s/2}^{-1},c_0,c_1^{s/2})=\sum\limits_{\forall c_i\in D}(r_i-1)^2, & c_i=c_0,\\ E(\theta,c_{-s/2}^{-1},c_0,c_1^{s/2})=\sum\limits_{\forall c_i\in D}(r_i-0)^2, & c_i\neq c_0,\end{cases} \tag{9}$$

其中，$r_i=l_2(y_i)=P_c(c_i=c_0\mid\theta,c_{-s/2}^{-1},c_1^{s/2},c_i)$.

在借鉴标准的 DAE 模型 stacking 策略[18]的基础上，我们构建出用于有色金属实体识别的 DNN 模型. 有色金属实体识别的深度学习模型 stacking 过程：

步骤 1. 利用训练只有单隐层的神经语言模型的方法来训练 1 个单隐层的文本窗口 DAE 网络；

步骤 2. 移除 softmax 输出层，增加另外一个隐层作为一个新的 softmax 输出层；

步骤 3. 按照步骤 1 的方法训练这个带有 2 个隐层的文本窗口 DAE 网络,这样就构建出一个深层的网络模型.

通过不断迭代这个 stacking 过程,直到堆叠出 1 个具有多个隐层的深度神经模型.

3.2 有色金属实体识别

在整个训练集上通过最大化 likelihood 来训练我们的神经网络,实现有色金属实体识别任务. 假定 θ 是网络中所有可训练的参数,T 表示训练的文本窗口数据集,模型训练的目标是使以下 log-likelihood 最大化:

$$\theta \mapsto \sum_{(x,y)\in T} \log p(y \mid x,\theta), \tag{10}$$

其中,x 表示当前正在训练的中文字符窗口及其关联的特征,y 代表相应的标记(如 B-M_PRO),$p(\cdot)$ 表示文本窗口中要预测的中心字符的标记的概率. 解决这个问题,目前有 2 种方法:

1) 基于词级别的 log-likelihood

这种方法是将文本窗口中的每个词看成是独立的,对于文本窗口中的一个输入字符 x,模型的输出是该字符$[f_\theta(x)]_i$ 的第 i 个标记(每个字符有多种可能的标记,如有色金属产品实体第 1 个字符 B-M_PRO 或有色金属实体内部字符 I-M_PRO)的打分,这个打分可以描述成一个带标记的条件概率 $p(i|x,\theta)$,是通过应用一个 softmax 函数[13]来计算所有标记.

$$p(i \mid x,\theta) = \frac{\mathrm{e}^{[f_\theta(x)]_i}}{\sum_j \mathrm{e}^{[f_\theta(x)]_j}}. \tag{11}$$

log-add 算子为

$$\log \underset{i}{\mathrm{add}}\, z_i = \ln\left(\sum_i \mathrm{e}^{z_i}\right), \tag{12}$$

式(12)表示当前词的相邻词的标记. 为此,把一个训练样例(x,y)的 log-likelihood 形式化表达描述为

$$\log p(y \mid x,\theta) = [f_\theta(x)]_y - \log \underset{j}{\mathrm{add}}[f_\theta(x)]_j, \tag{13}$$

式(13)表明在一个文本窗口(可能是一个句子)中,当前词的标记$[f_\theta(x)]_y$ 与它相邻词的标记$\log \underset{j}{\mathrm{add}}$ $[f_\theta(x)]_j$ 有紧密关联.

2) 基于句子级别的 log-likelihood

由于在一个句子或一个文本窗口中每个词语都具有非常强的上下文信息,针对我们的有色金属实体识别任务,预测中心字符在有色金属实体中的所有可能标记时,要充分考虑当前词与上下文中相邻词语标记之间的相关性. 为此,我们的模型在基于词级别的 log-likelihood 方法的基础上,充分考虑了相邻词语之间的标记信息,能有效提高对中心字符标记的打分. 我们考虑模型输出节点 $f_\theta([x]_1^T)$的打分矩阵,矩阵中的元素$[f_\theta(x)]_{i,t}$为模型输出节点的打分,表示文本窗口$[x]_1^T$ 中第 t 个词的第 i 个标记的打分. 在连续的词语中,一个词从标记 i 变换到标记 j 的过渡打分$[\boldsymbol{A}]_{i,j}$初始分值为$[\boldsymbol{A}]_{i,0}$,表示从第 i 个标记开始变换. 当训练这个过渡评分时,沿着标记 $[i]_1^T$ 路径的整个文本窗口的分值将通过对过渡打分及网络打分求和得到:

$$s([x]_1^T,[i]_1^T,\tilde{\theta}) = \sum_{t=1}^{T}([\boldsymbol{A}]_{[i]_{t-1},[i]_t} + [f_\theta(x)]_{[i]_t,t}), \tag{14}$$

其中,$\tilde{\theta}=\theta\cup\{[\boldsymbol{A}]_{i,j},\forall i,j\}$.

使用式(11)的 softmax 函数进行归一化文本窗口中所有输出节点的标记,并对所有可能的标记路径$[j]_1^T$ 进行归一化,把产生的比例作为标记路径的一个条件概率. 取对数,最终路径$[y]_1^T$ 的条件概率通过下式推理得到:

$$\log p([y]_1^T \mid [x]_1^T,\tilde{\theta}) = s([x]_1^T,[y]_1^T,\tilde{\theta}) - \log \underset{\forall [j]_1^T}{\mathrm{add}}\, s([x]_1^T,[j]_1^T,\tilde{\theta}), \tag{15}$$

式(15)中 log add 算子项的数目等于标记的数目,它会随着文本窗口的长度增长而膨胀. 为解决这个问题,可以根据下面的标准递归过程计算 log add 算子.

$$\begin{aligned}\delta_t(k) &\triangleq \log \underset{\{[j]_1^t \cap [j]_t = k\}}{\mathrm{add}}\, s([x]_1^t,[j]_1^t,\tilde{\theta}) = \\ &\log \underset{i}{\mathrm{add}} \log \underset{\{[j]_1^t \cap [j]_{t-1} = i \cap [j]_t = k\}}{\mathrm{add}} (s([x]_1^t,[j]_1^{t-1},\tilde{\theta}) + \\ &[\boldsymbol{A}]_{[j]_{t-1},k}) + [f_\theta(x)]_{k,t} = \\ &\log \underset{i}{\mathrm{add}}(\delta_{t-1}(i) + [\boldsymbol{A}]_{i,k}) + [f_\theta(x)]_{k,t},\ \forall k,\end{aligned} \tag{16}$$

按照以下条件终止:

$$\log \underset{\forall [j]_1^T}{\mathrm{add}}\, s([x]_1^T,[j]_1^T,\tilde{\theta}) = \log \underset{i}{\mathrm{add}}\, \delta_T(i). \tag{17}$$

为此,可以在整个训练数据集$([x]_1^T,[y]_1^T)$上使式(15)的 log-likelihood 最大化.

在推理时,给定一个要标记有色金属实体的文本窗口$[x]_1^T$,要使得式(14)中整个文本窗口的打分最小化,即必须要找到:

$$\arg\max_{j_1^T} s([x]_1^T,[i]_1^T,\tilde{\theta}). \tag{18}$$

使用 Viterbi 算法在对式(16)(17)递归执行时,首先将 log add 替换成 max;然后,通过每个 max 寻找出最优的标记路径,就可以实现有色金属实体识别.

4 实 验

为验证本文提出的模型对有色金属领域实体识别效果,在实验部分分别设置了5个对比实验:

实验1. 对比了本文提出的有色金属实体识别模型 DNN^3-CE_Y 与传统的BP神经网络模型NN(BP)及CRFs模型.

实验2. 验证了以中文字符构造的embeddings向量比以中文词语构造的embeddings向量对有色金属领域实体识别的效果好.

实验3. 对比了具有不同隐层数的DNN模型对识别结果的影响.

实验4. 对比了隐层中节点单元个数对实体识别效果的影响.

实验5. 对比了本文提出的深度学习模型分别对有色金属产品、有色金属组织机构、有色金属矿产、有色金属矿产地名这4类实体的识别效果.

其中,DNN^3 表示有3个隐层的深度神经网络模型;CE表示以字符的embeddings向量作为DNN模型的输入;WE表示以中文词语的embeddings向量作为DNN模型的输入;下标Y表示构造的embeddings向量融入了有色金属领域知识库的特征,如利用余弦值计算中文字符与相邻字符的距离时,结合了领域知网来计算.

另外,除实验5分别计算有色金属领域各类实体的识别效果外,其他实验中给出的准确率(precision, P)、召回率(recall, R)及 F_1 值都是指各类实体识别效果的平均值.

实验中我们严格按照标准评价指标,统计了各种方法的 P 和 R 两个指标,在这2个指标的基础上,利用 F_1 值作为衡量所提方法的最终评测指标. P,R 以及 F_1 值的公式如下:

$$P=\frac{\text{正确识别的有色金属实体个数}}{\text{识别的有色金属实体个数}}\times 100\%, \tag{19}$$

$$R=\frac{\text{正确识别的有色金属实体个数}}{\text{有色金属实体总数}}\times 100\%, \tag{20}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\%. \tag{21}$$

4.1 实验数据准备

实验数据是从百度百科、中文wiki百科、中国有色金属网、东方五金网、阿里巴巴冶金产品交易网、中国冶金产品信息网有色金属企业网站上收集有色金属产品相关页面,收集到的各类实体数量见表1所示,并将收集的文本进行标记.整个数据集分成2份,用来做模型训练的训练集包含40 046个有色金属领域实体,测试集包含1 344个有色金属实体数据.

Table 1 Entity Types in the Field of Nonferrous Metal

表1 有色金属领域实体类别

Serial Number	Entity Type of Nonferrous Metal Domain	Total Entity Number	Training Set	Test Set
1	Products of Nonferrous Metal	25 150	24 742	408
2	Organizations of Nonferrous Metal	1 270	983	287
3	Mines of Nonferrous Metal	14 280	13 741	539
4	Mineral Placenames of Nonferrous Metal	690	580	110

4.2 实验结果分析

实验1. 不同模型对有色金属领域实体识别结果的比较.

为了进行实验,使用Matlab中的神经网络工具包(neural network toolbox)来做BP实验,使用CRF++工具包来进行条件随机场模型的实验,表2分别给出了 DNN^3-CE_Y 模型、NN(BP)模型及CRFs模型对表1中所有实体的整体识别效果.NN(BP)模型及CRFs模型的输入是分词后的标注语料,特征是由人工方式选择的,特征选择是考虑了当前词、当前词的前后2个次的词性、IOBES标记等特征.

Table 2 Performance of Entity Recognition Based on Different Models in the Field of Nonferrous Metal

表2 不同模型识别有色金属领域实体结果比较 %

Model	P	R	F_1
NN(BP)	79.45	84.25	81.78
CRFs	80.37	84.94	82.59
DNN^3-CE_Y	81.29	86.21	83.68

从表2可以看出,传统的BP神经网络识别效果相对较差;CRFs模型识别效果比BP神经网络的效果高出0.81个百分点;DNN^3-CE_Y 模型效果最好,比CRFs模型的识别效果高出1.09个百分点.可见,基于本文提出的这种深架构的神经网络模型比传统的浅层模型效果好.

实验2. 分别以字和词构造的embeddings向量

对 DNN 模型识别效果的比较.

为了证明处理中文实体识别以字特征作为原始输入能有效避免因中文分词错误而影响到最终识别效果,本实验分别比较了以中文字符构造的 embeddings 向量和以中文词语构造的 embeddings 向量作为 DNN 模型的输入进行实体识别的效果.在构造字或词的 embeddings 向量时,文献[13]证明了任意一个词与它最临近的 10 个词的词法、句法、语义上具有非常强的关联关系,因此,可以用临近的 10 个词的特征来表征当前词.然而,在计算与当前词最临近的 10 个词的距离时,文献[13]是针对通用领域的计算方法,本文的任务是识别有色金属领域文本中的有色金属实体,有色金属实体中的字符间的紧密关系是通过字符之间的距离来计算的,因此,可以借助有色金属领域知识库来计算实体中字符之间的距离.另外,针对实体识别的任务,在构造 embeddings 向量时除了通用的特征外,还单独增加了当前字符及相邻字符是否在有色金属库中的特征,实验结果如表 3 所示:

Table 3 Influence of Chinese Character and Word Embeddings on the Recognition Performance

表 3 分别以字、词构造 embeddings 向量对识别性能的影响

%

Model	P	R	F_1
DNN^3-WE	79.68	84.40	81.97
DNN^3-CE	81.35	85.32	83.29
DNN^3-WE_Y	80.29	84.96	82.56
DNN^3-CE_Y	81.75	85.70	83.68

从表 3 可以看出,以字特征(DNN^3-CE)来构造 embeddings 向量比以词特征(DNN^3-WE)构造的 embeddings 向量能有效提高系统的识别效果;另外,增加了有色金属领域特征(DNN^3-CE_Y,DNN^3-WE_Y)的识别效果比未增加领域特征(DNN^3-CE,DNN^3-WE)的模型识别效果分别提高了 0.39 和 0.59 个百分点.

实验 3. 带有不同隐层数的 DNN 模型对识别结果影响的比较.

实验 3 证明了 DNN 模型中具有不同的隐层数对实体识别效果的影响,实验结果如表 4 所示.从表 4 可以看出,随着 DNN 模型中隐层数目的增加,模型的识别效果在不断提高;当隐层数增加到 4 个隐层时,由于层数增多导致模型训练的时间较长,而效果提高不是很明显.由此可见,使用 3 个隐层模型就能达到较好的效果;同时说明了 DNN 模型通过预训练自动学习特征的过程中,使用多个隐层的特征变换相比单个隐层的神经网络模型能提取到更有效的特征.

Table 4 Performance of Entity Recognition with Different Numbers of DNN Hidden Layers

表 4 具有不同隐层数的 DNN 模型识别结果比较 %

Model	P	R	F_1
DNN^1-CE_Y	78.42	85.86	80.97
DNN^2-CE_Y	80.27	86.55	82.45
DNN^3-CE_Y	81.47	83.68	83.68
DNN^4-CE_Y	81.48	86.00	83.69

实验 4. 隐层节点单元个数对实体识别效果影响的比较.

实验 4 的目的是验证 DNN 深层模型中每个隐层的节点个数对有色金属领域实体识别效果的影响.隐层中保留多少个节点单元能达到最好的效果?为了方便实验,我们的 DNN 模型是一个单隐层的神经网络模型,隐层中节点单元数从 300 逐渐增加到 1 800 个节点,实验结果如表 5 所示:

Table 5 Performance of Entity Recognition with Different Node Unit Numbers in Hidden Layers

表 5 验证隐层中节点单元个数对实体识别效果的影响

%

Model	P	R	F_1
DNN^1-CE_Y+300U	77.68	82.19	79.87
DNN^1-CE_Y+600U	78.47	83.46	80.89
DNN^1-CE_Y+900U	78.59	83.50	80.97
DNN^1-CE_Y+1200U	78.36	83.61	80.90
DNN^1-CE_Y+1500U	77.69	82.13	79.85
DNN^1-CE_Y+1800U	77.46	82.31	79.81

从表 5 可以看出,当隐层中节点单元数从 300 个节点增加到 600,900 时,实验效果在不断提高;但从 900 增加到 1 200,1 500,1 800 时,实验效果在逐渐降低.由此可见,在我们提出的深度模型中,节点单元数相对于原始输入的节点数,不能增加太多,也不能减少太多,减少太多节点数可能会丢失一些特征信息,增加太多可能会导致噪音信息,而且节点数增加也会导致模型训练复杂度增大,影响整个模型的性能.

实验 5. 有色金属领域各类实体的识别结果的比较.

实验 1～4 验证了我们的深度模型有 3 个隐层,并且隐层节点数为 900 时实验效果最好.为此,本实验在这个条件下单独对每类实体的识别效果进行了比较,实验结果如表 6 所示:

Table 6 Performance of Entity Recognition of Different Types in the Field of Nonferrous Metal

表 6 有色金属领域各类实体识别结果 %

Entity Types of Nonferrous Metal	P	R	F_1
Products	76.31	80.38	78.29
Organizations	84.59	89.07	86.77
Mines	80.19	84.33	82.21
Mineral Placenames	85.47	89.52	87.45

从表 6 可以看出,有色金属矿产地名实体识别效果最好,F_1 值为 87.45%;有色金属组织机构实体识别效果次之,F_1 值为 86.77%;而有色金属产品实体识别效果最差,F_1 值为 78.29%.原因可能是有色金属矿产地名实体、有色金属组织机构等实体内部有明显的领域特征,因此识别效果好;而对于有色金属产品类实体,由于实体内部组成复杂,可能是简单的有色金属化合物实体,也可能是简单的有色金属化合物实体和型号、规格、品牌等组成的实体(如纳米氧化铝(VK-L05C)、华伦铝合金电缆等),这类实体识别难度相对较大.

5 结 论

针对有色金属领域产品名、组织机构名、矿产名、地名这 4 类实体识别任务面临分词准确率不高、缺乏大量已标注的训练样本等问题,提出了一种基于深度神经网络架构的有色金属领域实体识别模型,利用 embeddings 词向量表示方法对原始输入的每个中文字符进行向量化表示,有效解决了采用传统的 one hot 向量表示存在的问题;利用神经语言模型训练方法对隐层的文本窗口降噪自动编码器模型进行逐层预训练,有效解决了有色金属领域实体识别任务特征提取问题.实验结果表明,我们提出的模型其 F_1 值相比 CRFs 模型及 BP 神经网络模型的 F_1 值分别提高了 1.09 个百分点和 1.9 个百分点.

下一步工作将结合有色金属领域实体识别任务,研究词向量的预训练方法来提高模型的性能.

参 考 文 献

[1] Mikheev A, Moens M, Grover C. Named entity recognition without gazetteers [C] //Proc of the 9th Conf on European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 1999: 1-8

[2] Zhou Guodong, Su Jian. Named entity recognition using an HMM-based chunk tagger [C] //Proc of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2002: 473-480

[3] Finkel J R, Manning C D. Nested named entity recognition [C] //Proc of the 2009 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2009: 141-150

[4] Chieu H L, Ng H T. Named entity recognition: A maximum entropy approach using global information [C] //Proc of the 19th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2002: 1-7

[5] Yoshida K, Tsujii J. Reranking for biomedical named-entity recognition [C] //Proc of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Stroudsburg, PA: ACL, 2007: 209-216

[6] Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: An experimental study [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1524-1534

[7] Liu X, Zhang S, Wei F, et al. Recognizing named entities in tweets [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2011: 359-367

[8] Liu Feifan, Zhao Jun, Lü Bibo, et al. Study on product named entity recognition for business information extraction [J]. Journal of Chinese Information Processing, 2006, 20(1): 7-13 (in Chinese)
(刘非凡, 赵军, 吕碧波, 等. 面向商务信息抽取的产品命名实体识别研究[J]. 中文信息学报, 2006, 20(1): 7-13)

[9] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief networks [J]. Neural Computation, 2006, 18(7): 1527-1554

[10] Yu Dong, Li Deng, Wang Shizhen. Learning in the deep-structured conditional random fields [C] //Proc of NIPS Workshop. New York: Academy Press, 2009: 1-8

[11] Yu Kai, Jia Lei, Chen Yuqiang, et al. Deep learning: Yesterday, today, and tomorrow [J]. Journal of Computer Research and Development, 2013, 50(9): 1799-1804 (in Chinese)
(余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804)

[12] Liu Jianwei, Liu Yuan, Luo Xionglin. Research and development on Boltzmann machine [J]. Journal of Computer Research and Development, 2014, 51(1): 1-16 (in Chinese)
(刘建伟, 刘媛, 罗雄麟. 玻尔兹曼机研究进展[J]. 计算机研究与发展, 2014, 51(1): 1-16)

[13] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537

[14] Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese relation extraction based on deep belief nets [J]. Journal of Software, 2012, 23(10): 2572-2585 (in Chinese)
(陈宇, 郑德权, 赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报, 2012, 23(10): 2572-2585)

[15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [C/OL] //Proc of Workshop at ICLR. 2013 [2014-03-10]. http://arxiv.org/pdf/1301.3781.pdf

[16] Wu K, Gao Z, Peng C, et al. Text window denoising autoencoder: Building deep architecture for Chinese word segmentation [G] //Natural Language Processing and Chinese Computing. Berlin: Springer, 2013: 1-12

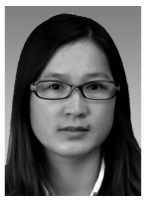
[17] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models [G] //Innovations in Machine Learning. Berlin: Springer, 2006: 137-186

[18] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research, 2010, 11(2): 3371-3408

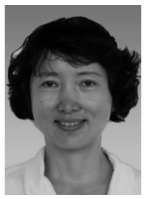
Mao Cunli, born in 1977. PhD and lecturer. Member of China Computer Federation. His main research interests include natural language processing, machine learning and information retrieval.

Yu Zhengtao, born in 1970. PhD, professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include natural language processing, machine learning and information retrieval.

Shen Tao, born in 1984. PhD, professor and PhD supervisor. Member of China Computer Federation. His main research interests include Terahertz technology, machine learning, information processing and non-destructive testing of materials.



Gao Shengxiang, born in 1977. PhD candidate. Student member of China Computer Federation. Her main research interests include natural language processing, machine translation and information retrieval (gaoshengxiang.yn@foxmail.com).



Guo Jianyi, born in 1964. Master and professor. Member of China Computer Federation. Her main research interests include natural language processing and information retrieval (gjade86@hotmail.com).

Xian Yantuan, born in 1981. PhD candidate. Student member of China Computer Federation. His main research interests include natural language processing and information retrieval (yantuan.xian@gmail.com).